Multimedia Tools and Applications (2024) 83:63919–63947
https://doi.org/10.1007/s11042-023-17991-2

Multimodal image registration techniques: a comprehensive survey

Henry O. Velesaca[1,2] · Gisel Bastidas[1,3] · Mohammad Rouhani[4] · Angel D. Sappa[1,5]

Received: 8 August 2023 / Revised: 14 December 2023 / Accepted: 24 December 2023 /
Published online: 6 January 2024

Abstract
This manuscript presents a review of state-of-the-art techniques proposed in the literature for multimodal image registration, addressing instances where images from different modalities need to be precisely aligned in the same reference system. This scenario arises when the images to be registered come from different modalities, among the visible and thermal spectral bands, 3D-RGB, or flash-no flash, or NIR-visible. The review spans different techniques from classical approaches to more modern ones based on deep learning, aiming to highlight the particularities required at each step in the registration pipeline when dealing with multimodal images. It is noteworthy that medical images are excluded from this review due to their specific characteristics, including the use of both active and passive sensors or the non-rigid nature of the body contained in the image.

Keywords Multimodal registration · Cross-spectral registration · Classical registration · Deep learning registration · Non-medical registration

✉ Henry O. Velesaca
hvelesac@espol.edu.ec ; hvelesaca@correo.ugr.es

Gisel Bastidas
giskbast@espol.edu.ec

Mohammad Rouhani
mohammad.rouhani@inria.fr

Angel D. Sappa
asappa@espol.edu.ec ; asappa@cvc.uab.es

1 FIEC, CIDIS, ESPOL Polytechnic University, Campus Gustavo Galindo, Guayaquil 09-01-5863, Ecuador

2 Software Engineering Department, University of Granada, Granada 18014, Spain

3 Escuela Superior Politécnica de Chimborazo, Panamerica Sur km 1 1/2, Riobamba, Ecuador

4 Inria Paris Centre, 2 Rue Simone Iff, Paris 75012, France

5 Computer Vision Center, Universitat Autónom de Barcelona, Bellaterra, Barcelona 08193, Spain

1 Introduction

The fields of image processing and computer vision are evolving towards scenarios where the coexistence of information from different sources, each contributing to efficient task-solving, is becoming more common. Some examples of such scenarios include video surveillance, combining images from the visible domain with 3D data from LiDAR or thermal imaging [1, 2]; the inspection of thermal insulation in buildings, utilizing visible and thermal images [3, 4]; environmental monitoring, where visible and NIR images are widely used in remote sensing applications [5]; image filtering and fusion, involving the combination of images from different sources for noise filtering and enhancement [6, 7]; and crop inspection, such as the detection of pests and diseases using images from various spectral bands acquired with UAVs or remote sensing [8]. Furthermore, in driving assistance systems, multimodal imaging is employed for pedestrian detection and scene understanding [9, 10]. These applications have been enabled by advancements in hardware technology, including modern smartphones equipped with multiple cameras that provide a rich source of information.

It should be highlighted that the success of the applications mentioned above is warranted if the different sources of information are correctly placed in the same reference system—this is referred to in the literature as an image registration problem. Image registration has been largely studied over the last 3 decades and many solutions have been proposed in the literature, mainly for registering images from the same domain acquired with the same sensor. In recent years, with the development of technology, the need to register images of a given scenario, acquired with sensors of different nature, has transformed the registration problem into a multimodal problem. The multimodal registration problem has mainly been studied in the framework of medical images, which is neither mentioned in the list of applications given above nor included in this survey. Medical image registration possesses unique characteristics, such as the use of active and passive sensors, or the need to handle non-rigid body structures. These specific challenges deserve separate attention and consideration in dedicated research efforts. The current survey focuses on the non-medical multimodal image registration problem, reviewing classical and novel deep learning-based approaches. Figure 1 illustrates some of the state-of-the-art applications reviewed.

Image registration is a process in which two or more images of the same scene, taken at different times or from different points of view, are aligned in order to combine or compare them. The process typically involves identifying matching features in the images and then applying a transformation to align them. The problem turns challenging when images from different modalities are considered, since features may look different due to the different nature of the sources. Hence, the challenge is to describe features in images from different modalities in a way that allows for unequivocal association. During last decades different hand-made feature descriptors have been proposed in the literature (e.g., [17–20]) to tackle the multimodal problem, most of them are adaptations of the well known EOH, SIFT, SURF, etc. (e.g., [21–23]). In recent years, deep learning-based methods have been proposed for multimodal image registration, which has shown promising results. These methods use Convolutional Neural Networks (CNNs) to learn image features (e.g., [24, 25]) used as a guidance to align the image during the registration process. The advantage of using CNNs is that they can learn to extract features from images even when the images are corrupted by noise or artifacts, or they correspond to different modalities.

This manuscript reviews the state-of-the-art multimodal image registration techniques proposed for the non-medical image domain. Firstly, a general image registration framework is presented detailing each of the steps present in any multimodal image registration process.

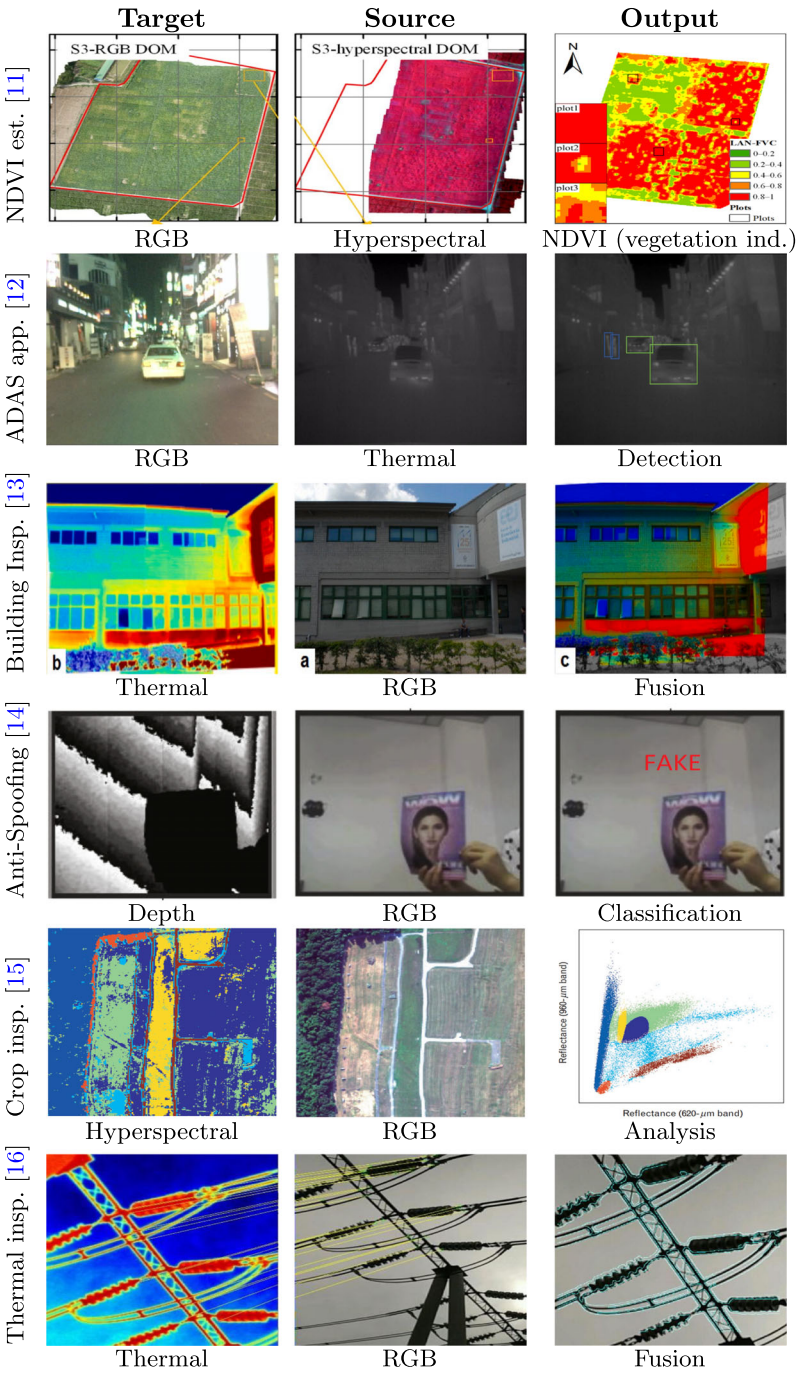

Fig. 1 Applications in the state of the art reviewed in the literature

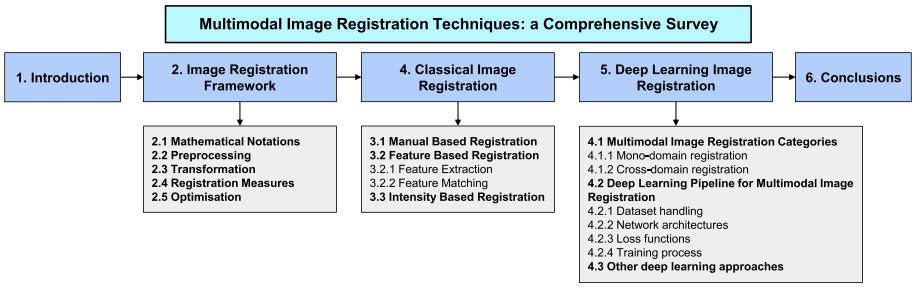

Fig. 2 Structure of this survey

Mathematical notation is introduced together with metrics used for the evaluation. Then, classical approaches are reviewed, classifying the proposed solutions according with the strategy used for the registration. Next, deep learning-based approaches proposed in the literature to tackle the multimodal registration problem are detailed. The main elements of deep learning approaches, such as dataset handling, network architectures, loss function and training process are reviewed. The manuscript ends with general conclusions summarizing the open problems and main challenges that still remain in multimodal image registration. Figure 2 shows the general structure of this work.

2 Image registration framework

This section starts by describing the mathematical notation used throughout the manuscript. Then, we review the most common image pre-processing approaches, focusing on enhancing both source and target images to improve the registration results. Next, transformation models usually applied in the registration of multimodal images are detailed. Then, registration measures proposed in the literature for the multimodal case are presented. Finally, optimization techniques generally used in this framework are just mentioned, since no particular requirement is needed for the multimodal case.

2.1 Mathematical notations

Table 1 show the mathematical notations used in the present work.

2.2 Preprocessing

An initial step in multimodal image registration involves preprocessing the source and target images. It includes a set of operations that are performed on the images to improve the registration results. The goal of preprocessing is to make the source and target images more similar and to reduce the variability between them, which can make the registration process more robust and accurate. It is worth mentioning that this step is optional and may or may not be present. There are a large number of techniques used in the preprocessing task of multimodal image registration, a couple of which are presented in this section together with some examples.

Table 1 Mathematical notations used in this paper

Symbol	Description
X	fixed image, also referred to as target or reference image
Y	moving image, also referred to as source image
$p(x, y)$	joint probability distribution of the intensities in images X and Y
$p(x)$	marginal probability distributions of the intensities in image X
$p(y)$	marginal probability distributions of the intensities in image Y
$MI(X, Y)$	mutual information between images X and Y
$H(X)$	entropy of image X
$H(Y)$	entropy of image Y
$x(i, j)$	intensity values at a point (i, j) in image X
$y(i, j)$	intensity values at a point (i, j) in image Y
$mean(X)$	mean intensity values of image X
$mean(Y)$	mean intensity values of image Y
$std(X)$	standard deviation of the intensity values of image X
$std(Y)$	standard deviation of the intensity values of image Y
N	total number of sample points in images X and Y
$\|X \cap Y\|$	area of overlap between images X and Y
$\|X\|$	number of pixels on image X
$\|Y\|$	number of pixels on image Y

One of the techniques used in preprocessing is **intensity normalization**, which involves normalizing the intensity values of the images to a common range, which can improve the registration results. This step can be performed using techniques such as histogram matching or intensity standardization. For instance, Okorie et al. [26] employ an intensity mapping technique that aims to transform the histogram of an image to match that of a reference image. This aims to produce a similar distribution of intensities in the image pair to be registered, and the method is specifically tailored for registering satellite images captured in different spectral bands.

The image **denoising** could be also considered as a preprocessing technique, it involves removing noise from the images for improving the registration results. This step can be performed using techniques such as median filtering, Gaussian filtering, or wavelet denoising. In Jiang et al. [27] the authors use Gaussian filtering to obtain the true pixel intensity and delete outliers progressively in order to improve the matching results. Registration of images from different spectral bands is considered. **Image enhancement** has been considered as another preprocessing technique. It is intended to enhance the given images in order to make the features more distinct for a better registration result. This step can be performed using techniques such as contrast stretching, histogram equalization, or unsharp masking. For example, Teke et al. [28] propose to use contrast stretching and histogram equalization to improve the correct match rate of the SURF method. According to the authors, better results are obtained with histogram equalization.

Contrary to all previous approaches, which are based on classical image preprocessing techniques, deep learning-based image-to-image translation has recently been used to transform the multimodal registration problem into a monomodal one, which is easier to solve. In [29], an empirical evaluation study for rigid registration of 2D and 3D images is performed.

The effectiveness of this strategy is evaluated by performing a GAN-based image-to-image translation method (e.g., pix2pix [30], CycleGAN [31], DRIT++ [32], or StarGANv2 [33]) or one contrastive representation learning model (CoMIR) [34] before the monomodal registration process. The results demonstrate that the use of the image-to-image translation method to preprocess the images mapping them in the same domain can be effectively performed prior to the monomodal registration technique in scenarios such as remote sensing; since in this kind of problem, the images will have high similarity in structure and appearance.

2.3 Transformation

Within the image registration process, the model transformation refers to the way the source (moving) image Y is aligned towards the target (reference) image X, different transformation models can be applied. The transformation model can be 2D, 3D, or other dimensions, depending on the dimensionality of the images. These models determine how the source image should be deformed for optimal alignment. From an optimization perspective, a transformation model defines the parameter space in which the optimal parameters are iteratively or simultaneously determined. The primary types of transformation models include rigid, affine, and non-rigid deformation. A rigid model allows only basic rotation and translation of the source image, while a non-rigid model offers more flexibility by allowing the source image to undergo complex deformations with higher degrees of freedom. In this section, we will explore and discuss the commonly used transformation models in multimodal image registration.

The first transformation model reviewed in this section is the **rigid registration**; this representation defines a rigid transformation, which includes translation and rotation, to align the images. This method is fast and simple, but it can be sensitive to changes in viewpoint, scale, and rotation. This simple transformation model is widely used in remote sensing applications, where images from different spectral bands are registered in the same reference system for further analysis (e.g., crop inspection ([35, 36]); furthermore, this transformation can be used in chemical imaging [37], face detection [38], or other applications ([39–41]), etc.). For instance, the work presented by [42] adopts a rigid transformation model in the registration of multimodal remote sensing images, and the authors use translations and rotation transformations to perform the alignment of images. Deng et al. [43] propose InMIR-Net, a network that can be used for both rigid and non-rigid multimodal image registration. For rigid registration, a homography matrix between the fixed and moving images is used in order to describe linear transformations such as scale, translation, rotation, and perspective. In addition, other deep learning-based approaches make use of a homography matrix for transformation [44–46].

Regarding the **affine registration**, it applies an affine transformation model that includes translation, rotation, scaling, and shearing, to align the images. This representation is more flexible than rigid registration, but it can still be sensitive to changes in viewpoint, scale, and rotation. In general, an affine transformation is composed of linear transformations (i.e., rotation, scaling, or shear) and a translation (or shift). Geometric contraction, expansion, dilation, reflection, rotation, shear, similarity, and translation are all affine transformations, as are their combinations. The authors in [47] propose a novel affine transformation model for registering cross-spectral images from satellites based on a robust transformation parameter estimation algorithm called the histogram of TAR sample consensus. In the same way, the authors in [48] present a novel affine and contrast invariant descriptor for infrared and visible image registration, in conjunction with using structural features extracted based on phase congruency. Additionally, Tu et al. [49] use spatial affine transformation in order to capture

the spatial correlation between two modalities. To this end, a Spatial Transformer Network (STN) is used in [50], the authors predict a deformation field between the images by using a Multi-level Refinement Registration Network. In [51], the authors propose the use of a projective transformation between the corresponding points on the corners of the image.

Finally, the **non-rigid registration**, which is also referred to as **deformable registration**, estimates the parameters of a non-rigid transformation. This transformation corresponds to a deformation field that aligns the source with the target image. Non-rigid deformations are more flexible than rigid or affine transformations and can handle large geometric variations and changes in viewpoint, scale, and rotation. For example, Ma et al. [52] propose a regularized Gaussian fields criterion for image registration between visible and infrared face images. The main idea is to represent the images using the edge maps and align them by a robust criterion with a non-rigid model. Non-rigid transformation is also used by Rabatel et al. [53]. The authors use the Fourier-Mellin transform to register visible and near-infrared unmanned aerial vehicle images. On the other hand, some learning-based approaches are using non-rigid transformation. For instance, Arar et al. [54] propose the use of non-rigid spatial transformation by applying a non-linear dense deformation between the multimodal images. Thin-Plate Spline (TPS) and Free-Form Deformation (FFD) are among other types of non-rigid transformations [55].

2.4 Registration measures

As mentioned above, the objective of registration techniques is to find the optimal transformation that aligns the source image with the target image. Therefore, it is crucial to define and understand the criteria by which we evaluate the alignment quality, known as the *registration error*. The choice of registration measure greatly influences the outcome of the registration process. These measures play a vital role not only in assessing the quality of the final alignment but also throughout the optimization phase, where the best transformation parameters are being sought.

One of the most widely used measures for evaluating multimodal image registration is the **mutual information** (MI). The basic idea behind MI is to measure the amount of information that is shared between the two images being registered. The main benefit of this measure is that it is invariant to monotonic transformations of the image intensities, this means that it does not depend on the specific intensity values, but rather on the relationship between them. This fact is important since the content of the images may look quite different since we are tackling images from different modalities. The mutual information between the target and source images (X and Y) is defined as:

$$MI(X, Y) = \sum x \sum y \cdot p(x, y) \cdot \log\left(\frac{p(x, y)}{p(x)p(y)}\right) \tag{1}$$

The MI value ranges between 0 and the highest possible value, which is the entropy of one of the images. A high mutual information value indicates a high degree of similarity between the images, while a low value indicates a low degree of similarity. Mutual information is sensitive to noise and intensity variations, which can lead to a high mutual information value even if the images are not well aligned. This is done by comparing the joint probability distribution of the intensities in the two images to the product of the marginal probability distributions of the intensities in each image. The algorithm proposed by [56], which has been tested on optical, LiDAR, SAR, and depth maps shows good performance against complex non-linear radiometric differences and improves the state-of-the-art values of mutual information.

A variant of the MI measure is **normalized mutual information** (NMI). The main idea is similar to MI, which is to measure the amount of information that is shared between the two images being registered. However, NMI addresses one of the limitations of MI by normalizing the mutual information value to compensate for changes in image intensity. It corrects for the effect of changes in image intensity, which can occur due to variations in imaging conditions or scanner settings. By normalizing the mutual information value, NMI can provide a more robust evaluation of the registration results. Also, NMI is invariant to monotonic transformations of the image intensities, similar to mutual information. However, like mutual information, it can be sensitive to noise and intensity variations, which can lead to a high NMI value even if the images are not well aligned. A drawback of both, MI and NMI, is that they do not take into account the spatial relationship between pixels, so it can be less suitable for images with large geometric variations [57]. The NMI between images X and Y is defined as:

$$NMI(X, Y) = 2 \cdot \frac{MI(X, Y)}{H(X) + H(Y)} \tag{2}$$

The NMI value ranges between 0 and 1; a value of 1 indicates perfect alignment between the images, while a value of 0 indicates that the images are completely independent. The authors in [58] propose a filter method of feature selection based on normalized mutual information. The average normalized mutual information is proposed as a measure of redundancy among features, in order to reduce the bias of MI toward multivalued attributes and restrict its value to the interval [0, 1].

Due to the different nature and appearance of the images to be registered, classical difference-based measures (e.g., Cross-Correlation (CC), Mean Square Error (MSE), etc.) are rarely used in multimodal image registration. A few works on a family of these difference based approaches can be found in the registration of hyperspectral or multispectral images. In these particular cases, these measures can be used since the images to be registered are similar due to the short distance between the acquired spectral bands. For instance, in [59] the authors implement a method to recover subpixel rotation and translation accuracy from an airborne hyperspectral imaging system using a portable hyperspectral tunable imaging system. In this case, images are registered using the **Normalized Cross-Correlation** (NCC). The normalized cross-correlation between images X and Y is defined as:

$$NCC(X, Y) = \sum \frac{(x(i, j) - mean(X)) \cdot (y(i, j) - mean(Y))}{std(X) \cdot std(Y)} \tag{3}$$

The NCC ranges between -1 and 1. A value of 1 indicates perfect alignment between the images, while a value of -1 indicates miss-alignment and a value of 0 indicates no correlation [60]. The main characteristic of the NCC over cross-correlation is that NCC is less sensitive to linear changes in the amplitude of illumination in the two compared images. Also, it is sensitive to both small and large intensity variations, providing a detailed analysis of the intensity variations between the images, particularly in the case of normalized cross-correlation where it takes into account the mean and standard deviation of the images. Another benefit of using NCC is that it is relatively simple to compute [56, 61].

The **Root Mean Square Error** (RMSE) is another measure that may be used for evaluating the registration results in the particular cases of multispectral or hyperspectral image registration, as mentioned above, since in general images to be registered look similar. To assess the registration results, RMSE is used in the overlapped area between the target image and the transformed source image. The idea is to measure the difference between the intensity values of the two images at each point, and then take the square root of the average of the squared differences. The RMSE value ranges between 0 and infinity, where 0 indicates perfect

alignment between the images, and the greater the RMSE value, the greater the dissimilarity between the images. It is important to note that the RMSE value is unit dependent, it is usually applied to normalized images [48, 62, 63]. For instance, in [35] the authors propose a novel remote sensing image registration method based on phase congruency and spatial constraint, to register SAR images. To evaluate the proposal, different pairs of multi-spectral satellite images are used. The RMSE between images X and Y is defined as:

$$RMSE(X, Y) = \sqrt{\frac{1}{N} \cdot \sum (x(i, j) - y(i, j))^2} \tag{4}$$

The **correlation coefficient** (CCO) is also used as a measure to evaluate the particular case of multispectral and hyperspectral registration results. The basic idea behind the correlation coefficient is to measure the linear relationship between the intensity values of two images at each point. The values of CCO range between -1 and 1. A value of 1 indicates perfect alignment between the images, while a value of -1 indicates miss-alignment and a value of 0 indicates no correlation (e.g., [64, 65]). In [66] the authors propose an integrated approach for the registration and fusion of hyperspectral and multispectral images. The correlation coefficient (CCO) between images X and Y is defined as:

$$CCO(X, Y) = \frac{\sum (x(i, j) - mean(X)) \cdot (y(i, j) - mean(Y))}{std(X) \cdot std(Y)} \tag{5}$$

The last measure reviewed in this work is the **Dice similarity coefficient** (DSC); it is used for evaluating the registration results in segmentation-based registration methods. The intuition behind DSC is to measure the similarity between two segmented images (e.g., binary masks or categorical data). The DSC value ranges between 0 and 1. A value of 1 indicates perfect overlap between the sets, while a value of 0 indicates that the sets have no elements in common (for more details see [67, 68]). The main characteristic of this measure is that it is specifically designed for comparing binary or categorical data, such as segmented images. It takes into account both the number of true positive and true negative pixels and provides a measure of the similarity between the segmented regions. The DSC is relatively insensitive to the size of the regions being compared, which can be useful in cases where the regions have different sizes or shapes. It is important to note that the images may be acquired by different sensors from various perspectives. One of the main drawbacks of DSC is that it assumes that the binary or categorical data being compared are perfect, which is often not the case in practice [69, 70]. The DSC is defined as:

$$DSC = 2 \cdot \frac{|X \cap Y|}{|X| + |Y|} \tag{6}$$

2.5 Optimisation

After detecting the features and establishing correspondences, the transformation function is applied to the source image. During the optimization step, we seek the optimal transformation parameters, aiming to minimize the registration measure. In general, this process is iteratively applied until a satisfactory alignment is achieved. Various optimization methods can be applied (e.g., simulated annealing, particle swarm optimization, gradient descent, Levenberg-Marquardt algorithm, genetic algorithm, to name a few); however, their use is exactly the same as in classical image registration, where both, source and target, belong to the same domain. There is not a specific particularity related to the multimodal problem that requires

some adaptation, hence we refer to review optimization approaches used in classical image registration problems in case the reader is interested in it.

It's important to note that the registration process can be sensitive to initial conditions, and running the process with different initial conditions can enhance the likelihood of finding an optimal solution. The number of iterations needed depends on the specific characteristics of the images, the application requirements, and the available computational resources. Some registration methods such as Elastix [71], Imregister [72] and Advanced Normalization Tools (ANTs) [73], are iterative by nature, the registration is based on the displacement field, and the algorithm estimates this field iteratively.

It is important to monitor the registration process during iteration, checking for convergence and ensuring that the registration results are not overfitting. Overfitting occurs when the registration process becomes too sensitive to noise and outliers, resulting in poor registration outcomes. One way to detect overfitting is by monitoring the change in registration metric values between iterations. Additionally, implementing early stopping criteria is crucial. This allows for terminating the registration process when the registration metric values reach a satisfactory level or when the registration results remain unchanged between iterations. This helps prevent the algorithm from running unnecessary iterations, saving computational resources.

3 Classical image registration

This section summarizes the state-of-the-art of classical approaches proposed in the literature to register multimodal images. The reviewed techniques are grouped into four categories depending on the way that points from source and target are selected and matched: manual registration, feature-based, intensity-based, and area-based methods. The following provides a summary of these categories along with descriptions of the techniques employed in each. It is worth mentioning that each approach has its strengths and weaknesses, and the choice of approach depends on the specific application and the characteristics of the images being registered.

3.1 Manual registration

Despite the advances in the image processing field, multimodal image registration is still often performed manually by experts in some particular applications (e.g., registration of images in cartography). This *manual* process consists of selecting control points, also referred to as landmarks, that will be used to establish correspondences between the images. The main idea is to manually or semi-automatically identify these points in both images and then use these points to align the images. The landmarks can be points, curves, or surfaces, depending on the nature of the structures being aligned [74, 75]. Manual registration can handle images with significant appearance differences, which may happen due to the geometric variations or the different nature of source and target images. These approaches are robust to changes in intensity, noise, and occlusions. Additionally, it can be useful when the region of interest is not well defined and the background is informative. The manual annotation requires significant expertise, time, and effort. It can also be subject to variability and errors due to individual differences between annotators. However, it can be a useful approach when the structures of interest are complex or subtle, and automated methods may not be accurate or reliable.

Manual annotation can provide a ground truth for evaluating the accuracy of automated registration methods.

Semi-automatic landmark registration refers to the process of selecting landmarks manually on one image and then automatically detecting corresponding landmarks in the other image to establish point correspondences. In this approach, the landmarks are manually identified by an expert, but their correspondence is determined automatically by a computer algorithm. This technique can be used in situations where the object of interest has easily identifiable landmarks that can be consistently located by the expert observer. Semi-automatic landmark registration methods can be quicker and more consistent than manual methods, yet retaining the accuracy of manually identified landmarks. The performance of these methods can be affected by the quality of the initial manual annotations, as well as the ability of the automated algorithm to correctly detect corresponding landmarks. For instance, Habib et al. [76] present a comprehensive investigation and implementation of the involved issues in a semi-automatic registration procedure capable of handling multi-source satellite imagery with varying geometric resolutions.

3.2 Feature based registration

As an enhancement of the previous section, various approaches have been proposed to automatically detect features in source and target images that are subsequently matched. To review the state-of-the-art approaches, this section is divided into two main subsections: 1. Describing feature points from different modalities (Section 3.2.1); and 2. Matching these feature points (Section 3.2.2).

3.2.1 Feature extraction

The term "feature" may refer to edges [77], corners [78], or blobs [79], which are employed to establish correspondences between pixels in the given source and target images. The fundamental idea behind such a representation is to extract a set of distinctive features from the images, and then use these features for image alignment. One advantage of feature-based registration is its ability to handle images with significant geometric variations and to be robust to changes in intensity, noise, and occlusions. Additionally, it can be useful when the region of interest is not well defined and the background is informative. Another benefit is its relative simplicity of computation, making it widely used in image processing. The literature proposes various handcrafted feature descriptors to achieve a similar representation of feature points in source and target images. Given that source and target images correspond to information from different modalities, this task is challenging. Furthermore, these descriptors should exhibit sensitivity to geometric variations [57, 80].

The description of points from different modalities cannot always be addressed with classical approaches (e.g., SIFT [81], SURF [23], RIFT [82], etc.) due to the non-linear intensity variations that may exist between images from the source and target. Hence, different strategies have been employed to achieve a comparable description, irrespective of the image modality. One notable contribution is presented in [83], where the authors propose a novel registration method, referred to as Uniform Robust SIFT (UR-SIFT), designed for various optical multi-source remote sensing images. The research addresses specific challenges in matching features in remote sensing images, considering scale variability and lighting conditions. Other early approaches focused on a straightforward strategy, which involves modifying gradient-based descriptors to operate within $[0, \pi]$ instead of $[0, 2\pi]$. This simple

modification helps to reduce the effect of changes in gradient direction between images from different modalities. It has been employed to adapt SIFT and HOG descriptors for matching visible and NIR images (e.g., [19, 20]). Although the strategy is simple, it significantly improves the performance of these algorithms in multimodal scenarios.

For the specific case of NIR and visible spectrum images, several approaches are based on the observations of [84], where a study of joint statistics between visible and thermal images is conducted. In this study, the authors identified a strong correlation between the boundaries of objects in both spectra. In other words, while texture information is lost, edge information remains similar between images from different spectral bands. Building on this discovery, [18] proposes a local Edge Histogram Descriptor (EHD) that utilizes edge information instead of image texture. Similarly, to address the LWIR and visible spectrum domain, the authors in [85, 86] suggest using EHD over image patches obtained by different Log-Gabor filters. More recently, Radhika et al. [87] extract point features from a pyramid of wavelet images, followed by a coarse-to-fine matching method based on a new aspect-based similarity measure. On the other hand, for RGB-D point cloud registration, Wan et al. [88] employ an enhanced ICP algorithm with infrared and color cameras.

Similar to the SIFT case, some authors propose modifications to the SURF descriptor to address the multimodal problem. For example, [89] proposes a method referred to as MM-SURF (Multimodal-SURF) that allows multimodality registration to be performed with high speed and accuracy for real-time applications. Taking the SURF algorithm as a starting point, Jhan et al. [90] propose to normalize this technique by naming it N-SURF, which can increase the number of correct matches between the pairs of images obtained from a multispectral camera and achieve one-step registration. The last method present by [82] is Radiation Variation Insensitive Feature Transform (RIFT). This method is a radiation insensitive image registration method based on phase congruency and a maximum index image. This method realizes the insensitivity and rotation invariance to multimodal image radiation changes. For instance, to address the problem of visible and thermal image registration-significant channels for remote sensing applications as mentioned in [91–94]-the authors in [95] use homomorphic filtering to enhance thermal images. They also employ a modified RIFT algorithm that utilizes a binary pattern string for descriptor construction.

3.2.2 Feature matching

Once feature points from both multimodal images are detected and described, the next step is to find correspondences between them that will be used to align the images and estimate registration parameters. In other words, once the features are described, the problem becomes similar to the feature matching in the monomodal case, since there is no particular consideration to be taken into account. The nearest key point is found by calculating the Euclidean distance between the corresponding descriptor vectors. Like in monomodal cases, matching robustness can be increased by using constraints such as the ratio between the first and second best matches should be higher than a given threshold [18]. Vijay et al. [96] propose an image registration system based on a local image feature descriptor matching algorithm. The algorithm is based on a Nearest Neighbour search, performed more efficiently using an acceptance threshold. The proposed feature-matching algorithm simultaneously matches n descriptors based on their total distance. This enabled a fast feature-based image registration. In addition, Yuan et al. [97] apply the NCC algorithm to achieve a one-to-one match between the UAV image feature points and the matching points on the corresponding Google satellite map. On the other hand, Song et al. [98] present a method to model and remove RGB-D cam-

era deviations using a specially designed checkerboard with hollow squares used to measure the deviations and camera parameters.

Due to the appearance difference on the image to be registered, a large number of wrong associations may appear during the feature matching, hence some filter to remove wrong matches is required like the well-known Random Sample Consensus (RANSAC) [99] algorithm largely used in computer vision. This method helps to estimate the parameters of a mathematical model from a set of observed data, which contains outliers. It randomly selects a minimal subset of feature correspondences and then estimates the parameters of the model using this subset. Subsequently, it evaluates the remaining correspondences to determine those that are inliers to the model. This method is robust to noise and outliers, but it can be sensitive to changes in viewpoint, scale, and rotation. For example, Cheng et al. [100] propose a new technical framework for remote sensing image matching by integrating affine invariant feature extraction and RANSAC. It is proposed to determine the distance threshold of RANSAC, which is a main problem in implementing this RANSAC-based automatic optimization. Finally, we should mention that optimal correspondences between feature points can be established through solving an assignment problem, using Hungarian algorithm, for instance [101].

3.3 Intensity based registration

The intensity values can be also used in some multimodal image registration approaches. This technique can be used when the images look similar, hence intensity values of the images may be used to establish correspondences between the images. The idea of this technique is to use the intensity values of the images to measure the similarity between the images at each point, and then use this similarity measure to align the images. This approach allows for a detailed analysis of intensity variations between the images. It is relatively simple to compute and finds wide applications in image processing. However, the main limitation is that it does not take into account the spatial relationship between the pixels, so it can be less suitable for images with large geometric variations. Additionally, it is sensitive to noise, so a high similarity measure can indicate not only a good registration but also a high noise level in the images.

Due to the limitations of the intensity-based approach, it is not possible to perform a correct alignment between visible-thermal images or RGB-depth images, since they present a large difference in intensity levels. Due to the aforementioned, Krishnan et al. [38] present an enhanced intensity-based image registration technique for thermal and visible facial images through a histogram matching scheme. Also, the result presented by the authors shows an improved registration quality when quantified using structural similarity and mutual information metrics. In a later study, Krishnan et al. [102] present an intensity-based image registration approach using visible and thermal face images. The proposed approach applies a saliency map strategy to balance the thermal and visible images intensity levels in order to overcome intensity differences and ensure proper multimodal image registration. This work presents good results using structural similarity index measure and universal image quality index metrics. In contrast, Chen et al. [103] propose an intensity-based multispectral image registration approach called normalized total gradient. In addition, Hu et al. [104] use an enhanced artificial bee colony algorithm with a leading group and MI metric to register multimodal images. The main idea of this measure is based on the assumption that the gradient of the difference between two aligned band images is sparser than that between two misaligned ones. The proposed approach consists of an image pyramid and global/local optimization;

it is introduced for affine transformation. The results show that this method is useful for unimodal and multimodal images.

Discussion

Classical approaches provide a viable solution for multimodal image registration when the source and target images are similar. However, in cases with significant appearance differences between images to be registered (e.g., visible spectrum and thermal; 3D and RGB), adaptations to the algorithms may be necessary to address this specific challenge. It is important to note that the pipeline presented in this manuscript can be modified depending on the characteristics of the images and the aim of the registration. Some steps could be skipped, or be combined, or even replaced by other methods. Additionally, the selection of each step will depend on the specific characteristics of the images, the requirements of the application, and the computational resources available. Table 2 summarizes the most recent classical approaches proposed in the literature for multimodal image registration.

4 Deep Learning based image registration

Classical image registration approaches are limited by their computational efficiency and the way these methods define the similarity measure metrics for the optimization process during the registration. To overcome the challenges of classical approaches, deep learning-based techniques are commonly used. These techniques measure similarity or estimate geometrical transformations. Depending on the training strategies, deep learning approaches are classified into supervised and unsupervised techniques, as demonstrated in studies such as [50, 54, 113].

There exist several deep learning-based architectures used for multimodal image registration; they typically involve Convolutional Neural Networks (CNNs), including ResNet, U-Net, and Fully Convolutional Networks (FCNs). Recently, transformers [44] and Generative Adversarial Networks (GANs) [46, 54] are also used during the registration process, as detailed in the next section. Most approaches use non-rigid or affine transformations (e.g., [44, 49, 114], [50, 54], to name a few). These deep learning-based techniques have demonstrated remarkable performance, significantly improving the accuracy and robustness of multimodal image registration. The manuscript delves into architectural innovations and training strategies that contribute to the success of these deep learning approaches.

4.1 Deep learning-based registration: categories

Deep learning-based multimodal image registration methods are commonly categorized into two groups: integrated learning methods [45, 51] and end-to-end learning methods [115]. Integrated learning methods involve incorporating a deep learning model into the classical multimodal image registration pipeline. These methods typically employ deep learning to estimate the similarity measurement between the source and target images, guiding the iterative optimization process. These techniques are also known as **deep iterative methods**. In contrast, end-to-end learning methods focus on directly estimating the transformation parameters and are referred to as **deep transformation estimation methods** [116]. Since end-to-end learning methods leverage deep learning techniques for the entire image registration process, various techniques have been proposed within this category.

Table 2 Summary of classical approaches proposed in the literature for multimodal image registration

Method	Image	Dataset	Evaluation metric	Code
A TIR-Visible Automatic Registration and Geometric Correction Method for SDGSAT-1 Thermal Infrared Image Based on Modified RIFT [95] **(2022)**				
Modified RIFT	Visible - TIR images	Landsat 8 and SRTM [105]	RMSE, Orthorectification acc.	☐
Intensity matching through saliency maps for thermal and visible image registration for face detection applications [102] **(2022)**				
Intensity matching through saliency maps	Visible - Thermal images	TUFTS [106, 107]	SSIM, Universal Image Quality index	☐
A generalized tool for accurate and efficient image registration of UAV multi-lens multispectral cameras by N-SURF matching [90] **(2021)**				
N-SURF	Multispectral images	Custom dataset [108]	Co-registration accuracy	☑
Histogram matched visible and infrared image registration for face detection [38] **(2021)**				
Histogram matched / SIFT and SURF	Visible - Infrared images	TUFTS [106, 107]	MI, SSIM	☐
Automated accurate registration method between UAV image and Google satellite map [97] **(2020)**				
RANSAC, NCC	UAV - Google satellite map images	Custom dataset	Positioning Error of the image	☐
Real-time adaptive visible and infrared image registration based on morphological gradient and C_SIFT [79] **(2020)**				
Morphological gradient and Improved SIFT	Visible - Infrared images	Custom dataset	-	☐
Remote sensing image registration based on phase congruency feature detection and spatial constraint matching [35] **(2018)**				
Phase congruency and spatial constraint	Remote sensing images		RMSE and Number of Correct Matches	☐
A novel affine and contrast invariant descriptor for infrared and visible image registration [48] **(2018)**				
Phase congruency	Visible - Infrared images	Custom dataset	Precision, Repeatability	☐

Table 2 continued

Method	Image	Dataset	Evaluation metric	Code
Feature based image registration using heuristic nearest neighbour search [96] **(2018)**				
Heuristic nearest neighbour search	Different viewpoints images	Oxford dataset [109]	RMSE	☐
Normalized Total Gradient: A new measure for multispectral image registration [103] **(2017)**				
Image pyramid and global/local optimization	Multispectral images	CAVE [110]	Normalized Total Gradient	☐
Robust registration of multimodal remote sensing images based on structural similarity [56] **(2017)**				
Histogram of orientated phase congruency	Visible - SAR - LIDAR - IR images, Map data	Custom dataset	SSIM	☐
LGHD: A feature descriptor for matching across non-linear intensity variations [86] **(2015)**				
Multi-scale and multi-oriented Log-Gabor filters	Visible - Depth - LWIR - NIR images	Custom dataset	Average Precision	☐
Rapid multimodality registration based on MM-SURF [89] **(2014)**				
Multimodal-SURF	Multimodality registration	Mikolajczyk's [111]	Average Error	☐
Multispectral stereo odometry [85] **(2014)**				
WBA incorporating Gauss-Newton optimization and RANSAC	Visible - Thermal images	Custom dataset	Error and RMSE	☐
Multispectral stereo image correspondence [77] **(2013)**				
Edges and Hough	Visible - LWIR images	FLIR ADAS [112]	-	☐

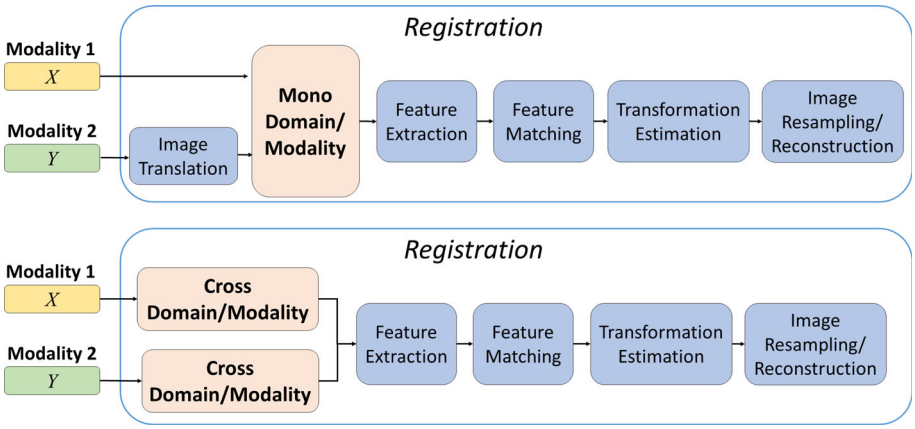

Fig. 3 Categories of end-to-end deep learning-based approaches

This paper provides a summary of state-of-the-art end-to-end deep learning approaches for multimodal image registration. These approaches are classified into two categories based on how multimodal images are translated for registration: mono-domain registration and cross-domain registration. The workflow of these categories is illustrated in Fig. 3.

4.1.1 Mono-domain registration

In this category, image translation is performed before estimating transformation parameters for registration. The registration process is executed using images represented in a mono-domain or mono-modality, involving feature extraction. These approaches utilize image-to-image translation techniques to convert multimodal image registration into monomodal image registration.

Arar et al. [54] propose a multimodal registration method that utilizes a spatial transformation network based on a fully convolutional network and a re-sampler layer for registration, a translation network to preserve geometric properties, and a discriminator based on a conditional GAN. The generator is used to produce a deformation field given two images. This method aims to learn a non-rigid spatial transformation for registration. In the registration network, the deformation field generator produces a matrix indicating the deformation direction for each pixel, representing the spatial transformation. These methods perform an image-to-image translation, where one of multimodal images is transformed into another modality. Pielawski et al. [34] propose a Contrastive Multimodal Image Representation (CoMIR) for registration, which allows for mono-domain registration since this method generates similar representations given two images of different modalities. CoMIR employs two U-Nets, each dedicated to one modality without weight sharing, along with a contrastive loss, InfoNCE.

"In addition, Wang et al. [50] present the Multi-Level Refinement Registration Network (MMRN) for the registration of infrared and visible images. Firstly, the visible image is translated to a pseudo-infrared image by the proposed Cross-Modality Perceptual Style Transfer Network (CPSTN); CPSTN is based on CycleGAN [31] but includes a perceptual style transfer constraint and cross regularizations across two-cycle learning paths. Then, the registration is performed in a mono-modality fashion by MMRN, predicting the deformation field between the infrared images and reconstructing the registered infrared image. Also in a mono-domain style, Xu et al. [113] propose a registration and fusion framework where the registration

module is performed through networks referred to as TransNet and AffineNet. TransNet performs a translation process to transfer the multimodal images into the same domain. Then, AffineNet generates the affine parameters, which are constrained by the similarity. Finally, Elsaeidy et al. [46] propose a keypoint-based Infrared-to-Optical image registration. In this approach, the infrared image is translated into an optical image, while keypoint extraction and matching are performed within a single domain.

4.1.2 Cross-domain registration

These approaches aim to extract features from each modality in order to register the multimodal images. Zhang et al. [51] propose a siamese fully convolutional network. Each branch is used for one modality and consists of seven convolutional layers. Tang et al. [115] propose a multiscale strategy to define a coarse-to-fine registration. In particular, three networks are cascaded on different scales. The first network uses Spatial Transformer Network which performs the initial registration. Then, the second network is used to generate a residual registration. The third network performs a deeper registration. To extract the semantic information and the geometric transformation parameters, the networks use Deep Residual ConvBlocks, a Squeeze Excitation, and Deep Residual ConvBlock with a channel attention mechanism. Besides, in [115] a structural similarity-based loss function is proposed to allow the registration of several multimodal images. Additionally, Tu et al. [49] propose a method that includes modality alignment for RGBT Salient Object Detection. The modality alignment module embeds a spatial transformer network.

In their work, Chen et al. [44] introduce a network known as Shape-Former, incorporating a robust soft estimation of outliers to filter outliers before extracting shape features. Additionally, they propose ShapeConv to integrate CNN with Transformer. A coarse registration layer is at the front end of Shape-Former for generating a homography matrix. Additionally, Quan et al. [117] propose a CNN multimodal feature learning and matching model, which is optimized by self-distillation learning, matching learning, and reconstruction learning. This model based on a partially unshared network allows extracting features from each modality in different branches. Then, a shared feature mapping model is used to map the features into a shared space before being matched. The self-distillation learning is used to extract richer similarity information while reconstruction learning helps to separate matching features from nonmatching features. Debaque et al. [114] propose a method for the registration of thermal and visible images using homography estimation to correct parallax from two images. This method is based on ResNet34 architecture, which consists of three modules: a feature extractor, a mask predictor, and a mapping computation.

Discussion

Deep iterative methods can be slow in highly dimensional space since the transformation estimation is iterative. Hence, end-to-end approaches can be used in a high-dimensional space. Furthermore, it is important to mention that the loss function plays a vital role in the deep learning process. Hence, several studies have proposed novel loss functions to improve registration keeping the structural similarity. Besides, a registration process may consist of multiple network architectures, where several loss functions can be used such as structural similarity, triplet loss, reconstruction loss, and L1-reconstruction loss, to name a few — loss functions are detailed in Sec. 4.2.3.

4.2 Deep learning-based registration: pipeline

Deep learning approaches can exhibit variations in their pipelines due to factors such as network architecture, choice of loss functions, data handling, and the training process. This section summarizes these factors of deep learning-based approaches for multimodal image registration. It is crucial to keep in mind that deep learning techniques are used to perform the steps of the classical multimodal image registration based on the learning process for feature extraction, matching, transformation estimation, and reconstruction.

4.2.1 Dataset handling

Data augmentation is a technique used to increase the size of the training dataset by applying random transformations to the images, such as rotation, scaling, shearing, and flipping. Given the limited availability of aligned data in multiple modalities, data augmentation can be employed to generate a larger dataset. Geometric transformations are typically used for data augmentation. For instance, Parbs et al. [118] use random flipping, rotation, and cropping as data augmentation techniques. Also, Gaussian noise is used for slightly deforming the input images. Tang et al. in [115] use data augmentation for training the registration network. The dataset is augmented through random affine transformations applied to aligned images to obtain four corner displacements. Additionally, Lu et al. [29] use random horizontal flip, rotation, random Gaussian blur, and crop.

Deep learning-based techniques, particularly Generative Adversarial Network (GAN), are also used for data augmentation. For instance, Quan et al. [119] propose the use of a GAN in order to generate additional data for optical and SAR image registration. The generator of the proposed GAN is optimized for the distribution, pixel, and reconstruction aspects of the images using adversarial, mapping, and reconstruction losses.

4.2.2 Network architectures

Network architecture refers to the structure of the neural network model, determining how the data flows through the network and how it is processed and transformed during the learning process. These methods can be supervised [44] or unsupervised (e.g., [50, 54, 113, 115]). The network architectures typically used for multimodal image registration are based on CNN, however, some approaches use other architectures such as transformers, and GANs. Figure 4 shows the general schemes for multimodal image registration based on the aforementioned deep learning architectures. These approaches are detailed as follows.

CNN architecture is composed of several blocks such as convolution layers, pooling layers, and fully connected layers. Quan et al. [117] present a partially unshared feature extraction network for optical-SAR images based on a Siamese network. This network consists of two low-level feature extraction networks one for each modality since there are notable differences in appearance between optical and SAR features. The multimodal feature learning model comprises two convolutional layers, followed by a batch normalization step. After feature extraction, the common features are mapped in a shared approach. The shared feature learning model includes five convolutional layers. On the other hand, Zhang et al. [51] propose a siamese fully convolutional network. This network contains seven convolutional layers, along with the ReLU activation layer and batch normalization layer.

In [115], the registration method extracts semantic information and generates geometric transformation parameters by utilizing a channel attention mechanism and a residual network

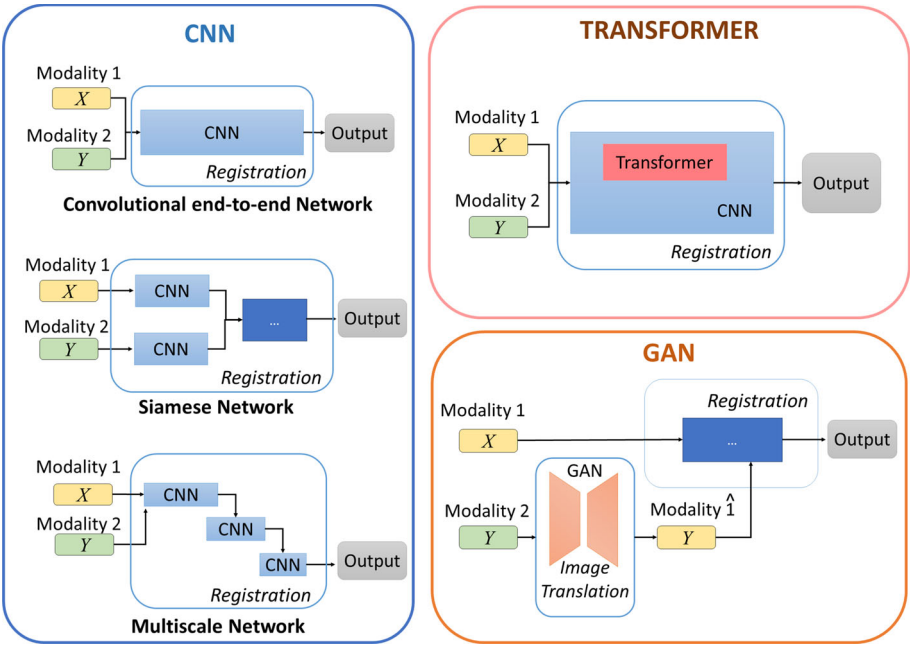

Fig. 4 Overview of architectures used for multimodal image registration

in a multiscale coarse-to-fine strategy. A residual network contains building blocks called residual blocks, which consist of two or more convolutional layers with batch normalization and non-linear activation functions with a skip connection. In [45] the authors use a CNN only for feature extraction. The backbone of this network is VGG-16.

The self-attention mechanism, introduced in **transformers**, captures relationships between different positions in a given sequence to compute a representation of the sequence [120]. The Transformer architecture consists of an encoder and a decoder using stacked self-attention and point-wise, fully connected layers. Chen et al. [44] propose the use of a CNN and a transformer for image registration. In particular, the transformer is used to improve the matching for good correspondences, that is, the network can learn long-range interactions from sparse matches in order to avoid outliers. Moreover, Tu et al. [49] propose the use of a Spatial Transformer Network (STN) to predict a transformation matrix. Hence, Transformers are used in the registration process as a part of the general framework to perform a specific task such as to generate a transformation matrix or improve the feature extraction.

GAN consists of a generator network and a discriminator network. The generator learns to produce synthetic data by mapping random noise or latent vectors to the desired data distribution. On the other hand, the discriminator distinguishes between real data samples and fake samples generated by the generator [121]. In the context of registration, this architecture is used for image translation in order to generate a domain-to-domain mapping. Pielawski et al. [34] use generative models such as pix2pix [30] and CycleGan [31] for image translation before registration. In addition, Elsaeidy et al. [46] propose to use a GAN to convert infrared images into optical images in order to perform a monomodal registration. The

generator backbone is a U-Net 256, while the discriminator is tested with two architectures: PatchGAN and PixelGAN. PatchGAN classifies individual patches and PixelGAN classifies pixels. Besides, Wang et al. [50] propose a cross-modality perceptual style transfer network for image translation, which is based on a GAN to generate pseudo infrared images.

4.2.3 Loss functions

The loss function is important for deep learning techniques during the learning process. Several studies have proposed different loss functions for improving image registration. Three loss functions are proposed in [117] for matching learning, self-distillation learning, and reconstruction learning. The proposed loss functions include the triplet loss, the feature consistency loss, and the reconstruction loss. The triplet loss aims to minimize the feature distance of matching image pairs. While the feature consistency loss focuses on the similarity of low-level and high-level features. The reconstruction loss constrains the matching features between modalities to be quite similar. Tang et al. [115] propose the use of structural similarity as a loss function for the end-to-end registration process. Since this process is performed in multiple scales, the overall loss is computed by summing the individual losses from each scale.

Zhang et al. [51] propose a loss function that aims to maximize the dissimilarity between the correct matched features and the nearby non-matching features. In [50], the authors propose the use of bidirectional similarity loss for the registration process and Multiscale Structural Similarity Index (MS-SSIM) loss for the fusion process. This aims to maintain the sharper intensity distribution of the fused image. Furthermore, Arar et al. [54] use L1-reconstruction loss and adversarial loss.

4.2.4 Training process

The next step is to train the network using a dataset of multimodal image pairs. The end-to-end networks learn to register the moving image to the fixed image during training. The training is commonly performed in one stage. However, some approaches perform the training in two stages. For instance, in [115], the coarse-to-fine registration network, which consists of three models stacked in a cascading way, is trained in two stages: initialization and joint training. In the initialization stage, the models are trained individually and successively to initialize the weights; that is, the first model is trained to obtain fixed weights for the second model in the stacked framework. Then, in the joint training, the overall network is trained with no fixed weights. On the other hand, some networks are not trained from scratch, that is, pre-trained subnetworks can be used to fine-tune the learning process.

Discussion

It is worth mentioning that registration techniques can be performed in 2D or 3D, and there are many variations in this pipeline depending on the specific task and modalities. The fundamental concept is to employ a deep learning model to learn the relationships between images and use that learned model to align different modalities. Typically, a CNN serves as a module for an end-to-end registration process. However, other architectures like Transformers and GANs are integrated into the registration process for specific tasks, such as image translation.

Table 3 Deep Learning-based approaches for multimodal image registration

Category	Method	Image	Dataset	Evaluation metric	Code
Shape-Former: Bridging CNN and Transformer via ShapeConv for multimodal image matching [44] **(2023)**					
Cross-domain	CNN and Transformer	Remote sensing and other computer vision multimodal images	Multimodal image matching datasets [116], and Remote sensing Dataset [124]	F-score, Precision, Recall	☐
Self distillation feature learning network for optical and sar image registration [117] **(2022)**					
Cross-domain	Siamese network	Optical - SAR images	SEN1-2	FPR95, Accuracy	☐
Unsupervised Misaligned Infrared and Visible Image Fusion via Cross Modality Image Generation and Registration [50] **(2022)**					
Mono-domain	CNN	Visible - Infrared images	TNO [125] and Road-Scene [126]	SSIM, Cross Correlation, Visual Information Fidelity	☑
RFNet: Unsupervised Network for Mutually Reinforcing Multi-modal Image Registration and Fusion [113] **(2022)**					
Mono-domain	CNN	Visible - Near Infrared images	MSIFT [127]	RMSE, SSIM, Max Square Error, Median Square Error and Peak Signal-to-Noise Ratio	☐
Infrared-to-Optical Image Translation for Keypoint-Based Image Registration [46] **(2022)**					
Mono-domain	GAN and Shift	Infrared - Optical images	Custom dataset	nn-mAP, Repeatability, Matching score	☐
Weakly Alignment-Free RGBT Salient Object Detection With Deep Correlation Network [49] **(2022)**					
Cross-domain	CNN (LSTM)	Visible - Thermal images	VT821 [128], VT1000 [129] and VT5000 [130]	F-measure, Weighted F-measure, Mean Absolute Error, S-measure, E-measure	☐
Thermal and Visible Image Registration Using Deep Homography [114] **(2022)**					

Table 3 continued

Category	Method	Image	Dataset	Evaluation metric	Code
Cross-domain	CNN	Visible - Thermal images	VLIRVDIF [131]	Average reprojection error between the ground truth predicted homologous points	☐
Multi-Source Remote Sensing Image Registration Based on Local Deep Learning Feature [45] (2021)					
Cross-domain	CNN	Optical - SAR images	Custom dataset	Pixel accuracy	☐
Unsupervised multi-modal image registration via geometry preserving image-to-image translation [54] (2020)					
Mono-domain	Spatial transformation network	Visible - Infrared images	Commercial dataset	NCC, SSIM	☐
Comir: Contrastive multimodal image representation for registration [34] (2020)					
Mono-domain	Contrastive learning	Visible - NIR images	Zurich dataset	MSE, RMSE, Cosine Similarity	☐
Registration of Multimodal Remote Sensing Image Based on Deep Fully Convolutional Neural Network [51] (2019)					
Cross-domain	Siamese fully convolutional network	Optical-NIR, Optical-TIR, Optical-SAR, Optical-Map	Google Earth historic images, WorldView-2, Landsat-8, TerraSAR-X	RMSE, Score Map, Correctly Matching Rate, Matching Precision	☐

4.3 Other deep learning approaches

Registration is an important task for other low-level vision tasks such as multimodal image fusion where the images from the same scene are required to be aligned. In [122], authors highlight the challenge of handling partial or globally unaligned multimodal images in image fusion frameworks. Therefore, recent approaches have included a registration module within the fusion framework. For instance, [44] emphasizes the importance of the registration process in a subsequent image fusion network. In particular, the U2Fusion network [123] is employed for fusion after the registration approach. Xu et al. [113] includes a registration module into a fusion framework, referred to as RFNet. Furthermore, Wang et al. [50] present a fusion framework that includes a Multi-Level Refinement Registration Network (MMRN) to register infrared and visible images.

Moreover, the registration process can be applied within vision application frameworks, as demonstrated in the RGBT Salient Object Detection framework proposed by [49]. This framework includes a Modality Alignment Module (MAM), comprising a spatial affine transformation component, a feature-wise affine transformation component, and a dynamic convolution layer component.

Discussion

Currently, there are a few approaches that combine multimodal registration and fusion. Thus, for future works, it will be important to consider integrating registration as a prerequisite procedure for fusion in a single deep learning-based framework, as they are complementary tasks that can improve the performance of various applications, such as remote sensing, security, autonomous driving, etc. Table 3 summarizes recent approaches for multimodal image registration based on deep learning architectures.

5 Conclusions

In this manuscript, we have undertaken an extensive exploration of multimodal image registration techniques, encompassing both classical methodologies and state-of-the-art deep learning-based solutions. By addressing the unique challenges and considerations inherent to multimodal images, our review offers a valuable resource to researchers and practitioners aiming to improve registration accuracy and broaden its applications across diverse domains. As technology continues to evolve towards multi-sensor platforms, this review also catalyzes the research community to conceive innovative solutions for the multimodal registration problem. The main features of reviewed papers are thoughtfully summarized in two tables, categorizing them according to their respective approaches: classical methodologies and deep learning-based strategies. These tables provide readers with a concise and organized overview, enabling them to quickly identify relevant studies and gain insights into the strengths and limitations of each approach. Overall, this manuscript not only consolidates the existing knowledge on multimodal image registration but also stimulates further advancements in this critical area of research, setting the stage for enhanced capabilities in image fusion and analysis.

Acknowledgements This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-22-1-0261; and partially supported by the Grant PID2021-128945NB-I00 funded

by MCIN/AEI/10.13039/501100011033 and by “ERDF A way of making Europe"; the “CERCA Programme / Generalitat de Catalunya"; and the ESPOL project CIDIS-12-2022.

Data Availibility Statement Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Ekpar F (2008) A framework for intelligent video surveillance, 421–426 (IEEE)
2. Torresan H, Turgeon B, Ibarra-Castanedo C, Hebert P, Maldague XP (2004) Advanced surveillance systems: combining video and thermal imagery for pedestrian detection, Vol. 5405, 506–515 (SPIE)
3. Yu X, Tian X (2022) A fault detection algorithm for pipeline insulation layer based on immune neural network. Int Journal of Pressure Vessels and Piping 196:104611
4. Kim C, Park G, Jang H, Kim E-J (2022) Automated classification of thermal defects in the building envelope using thermal and visible images. Quantitative InfraRed Thermography Journal 1–17
5. Asadzadeh S, de Oliveira WJ, de Souza Filho CR (2022) Uav-based remote sensing for the petroleum industry and environmental monitoring: State-of-the-art and perspectives. J Pet Sci Eng 208:109633
6. Li X, Ye H, Qiu S (2022) Cloud contaminated multispectral remote sensing image enhancement algorithm based on mobilenet. Remote Sensing 14:4815
7. Pan Y, Liu D, Wang L, Xing S, Benediktsson JA (2022) A multispectral and panchromatic images fusion method based on weighted mean curvature filter decomposition. Appl Sci 12:8767
8. Hafeez A et al. (2022) Implementation of drone technology for farm monitoring & pesticide spraying: A review. Information Processing in Agriculture
9. Lahmyed R, El Ansari M, Ellahyani A (2019) A new thermal infrared and visible spectrum images-based pedestrian detection system. Multimedia Tools Appl 78:15861–15885
10. Nam Y, Nam Y-C (2018) Vehicle classification based on images from visible light and thermal cameras. EURASIP Journal on Image and Video Processing 1–9
11. Yue J et al (2021) Method for accurate multi-growth-stage estimation of fractional vegetation cover using unmanned aerial vehicle remote sensing. Plant Methods 17:1–16
12. Hwang S, Park J, Kim N, Choi Y, So Kweon I (2015) Multispectral pedestrian detection: Benchmark dataset and baseline 1037–1045
13. Shariq MH, Hughes BR (2020) Revolutionising building inspection techniques to meet large-scale energy demands: A review of the state-of-the-art. Renew Sustain Energy Rev 130:109979
14. Jia Y, Zhang J, Shan S (2021) Dual-branch meta-learning network with distribution alignment for face anti-spoofing. Trans Inf Forensics Secur 17:138–151
15. Patel H, Upla KP (2020) Night vision surveillance: Object detection using thermal and visible images 1–6 (IEEE)
16. Cheng T, Gu J, Zhang X, Hua L, Zhao F (2022) Multimodal image registration for power equipment using clifford algebraic geometric invariance. Energy Rep 8:1078–1086
17. Yi Z, Zhiguo C, Yang X (2008) Multi-spectral remote image registration based on SIFT. Electron Lett 44:1
18. Aguilera C, Barrera F, Lumbreras F, Sappa AD, Toledo R (2012) Multispectral image feature points. Sensors 12:12661–12672
19. Pinggera12 P, Breckon T, Bischof H (2012) On cross-spectral stereo matching using dense gradient features 2:3
20. Firmenichy D, Brown M, Susstrunk S (2011) Multispectral interest points for RGB-NIR image registration 181–184 (IEEE)
21. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60:91–110
22. Vural MF, Yardimci Y, Temizel A (2009) Registration of multispectral satellite images with orientation-restricted SIFT, Vol. 3, III–243 (IEEE)
23. Bay H, Tuytelaars T, Van Gool L (2006) Surf: Speeded up robust features. Lect Notes Comput Sci 3951:404–417

24. Balntas V, Johns E, Tang L, Mikolajczyk K (2016) PN-Net: Conjoined triple deep network for learning local image descriptors. arXiv:1601.05030
25. Zagoruyko S, Komodakis N (2015) Learning to compare image patches via convolutional neural networks 4353–4361
26. Okorie A, Makrogiannis S (2019) Region-based image registration for remote sensing imagery. Comput Vision Image Underst 189:102825
27. Jiang X et al (2020) Robust feature matching for remote sensing image registration via linear adaptive filtering. Trans Geosci Remote Sens 59:1577–1591
28. Teke M, Temizel A (2010) Multi-spectral satellite image registration using scale-restricted surf 2310–2313 (IEEE)
29. Lu J, Öfverstedt J, Lindblad J, Sladoje N (2022) Is image-to-image translation the panacea for multimodal image registration? a comparative study. Plos one 17:e0276196
30. Isola P, Zhu J-Y, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks 1125–1134
31. Zhu J-Y, Park T, Isola P, Efros AA (2017 Unpaired image-to-image translation using cycle-consistent adversarial networks
32. Lee H-Y et al (2020) Drit++: Diverse image-to-image translation via disentangled representations. Int J Comput Vision 128:2402–2417
33. Choi Y, Uh Y, Yoo J, Ha J-W (2020) Stargan v2: Diverse image synthesis for multiple domains 8188–8197
34. Pielawski N et al (2020) Comir: Contrastive multimodal image representation for registration arXiv:2006.06325
35. Ma W, Wu Y, Liu S, Su Q, Zhong Y (2018) Remote sensing image registration based on phase congruency feature detection and spatial constraint matching. Access 6:77554–77567
36. Li K, Zhang Y, Zhang Z, Lai G (2019) A coarse-to-fine registration strategy for multi-sensor images with large resolution differences. Remote Sensing 11:470
37. Tarolli JG, Bloom A, Winograd N (2016) Multimodal image fusion with sims: Preprocessing with image registration. Biointerphases 11:02A311
38. Krishnan PT, Balasubramanian P, Jeyakumar V (2021) Histogram matched visible and infrared image registration for face detection 222–226 (IEEE)
39. Banharnsakun A, Achalakul T, Sirinaovakul B (2011) The best-so-far selection in artificial bee colony algorithm. Appl Soft Comput 11:2888–2901
40. Debayle J, Presles B (2016) Rigid image registration by general adaptive neighborhood matching. Pattern Recogn 55:45–57
41. Velesaca HO, Vulgarin J, Vintimilla BX (2023) Deep learning-based human height estimation from a stereo vision system 1–7
42. Yan X, Zhang Y, Zhang D, Hou N, Zhang B (2020) Registration of multimodal remote sensing images using transfer optimization. Geosci Remote Sens Lett 17:2060–2064
43. Deng X, Liu E, Li S, Duan Y, Xu M (2023) Interpretable multi-modal image registration network based on disentangled convolutional sparse coding. Trans Image Process 32:1078–1091
44. Chen J et al (2023) Shape-former: Bridging cnn and transformer via shapeconv for multimodal image matching. Inf Fusion 91:445–457
45. Zhang Y, Zhang Z, Ma G, Wu J (2021) Multi-source remote sensing image registration based on local deep learning feature. International Geoscience and Remote Sensing Symposium 2021-July, 3412–3415
46. Elsaeidy M, Erkol ME, Gunturk BK, Ates HF (2022) Infrared-to-optical image translation for keypoint-based image registration (Institute of Electrical and Electronics Engineers Inc.)
47. Song Z, Zhou S, Guan J (2013) A novel image registration algorithm for remote sensing under affine transformation. Trans Geosci Remote Sens 52:4895–4912
48. Liu X, Ai Y, Zhang J, Wang Z (2018) A novel affine and contrast invariant descriptor for infrared and visible image registration. Remote Sensing 10:658
49. Tu Z, Li Z, Li C, Tang J (2022) Weakly alignment-free rgbt salient object detection with deep correlation network. Trans Image Process 31:3752–3764
50. Wang D, Liu J, Fan X, Liu R. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration 3508–3515
51. Zhang H et al (2019) Registration of multimodal remote sensing image based on deep fully convolutional neural network. Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12:3028–3042
52. Ma J, Zhao J, Ma Y, Tian J (2015) Non-rigid visible and infrared face registration via regularized gaussian fields criterion. Pattern Recogn 48:772–784
53. Rabatel G, Labbe S (2016) Registration of visible and near infrared unmanned aerial vehicle images based on fourier-mellin transform. Precis Agric 17:564–587

54. Arar M, Ginger Y, Danon D, Bermano AH, Cohen-Or D (2020) Unsupervised multi-modal image registration via geometry preserving image-to-image translation 13410–13419
55. Rouhani M, Sappa AD (2012) Non-rigid shape registration: A single linear least squares framework, Vol. 7578, 264–277 (Springer)
56. Ye Y, Shan J, Bruzzone L, Shen L (2017) Robust registration of multimodal remote sensing images based on structural similarity. Trans Geosci Remote Sens 55:2941–2958
57. Zitova B, Flusser J (2003) Image registration methods: a survey. Image Vision Comput 21:977–1000
58. Estévez PA, Tesmer M, Perez CA, Zurada JM (2009) Normalized mutual information feature selection. Trans Neural Netw 20:189–201
59. Erives H, Fitzgerald GJ (2006) Automatic subpixel registration for a tunable hyperspectral imaging system. Geosci Remote Sens Lett 3:397–400
60. Zhao F, Huang Q, Gao W (2006) Image matching by normalized cross-correlation, Vol. 2, II–II (IEEE)
61. Rao YR, Prathapani N, Nagabhooshanam E (2014) Application of normalized cross correlation to image registration. Int J Res Eng Technol 3:12–16
62. Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)?-arguments against avoiding RMSE in the literature. Geosci Model Dev 7:1247–1250
63. Willmott CJ, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res 30:79–82
64. Asuero AG, Sayago A, González A (2006) The correlation coefficient: An overview. Critical Reviews in Analytical Chemistry 36:41–59
65. Taylor R (1990) Interpretation of the correlation coefficient: a basic review. Journal of Diagnostic Medical Sonography 6:35–39
66. Zhou Y, Rangarajan A, Gader PD (2019) An integrated approach to registration and fusion of hyperspectral and multispectral images. Trans Geosci Remote Sens 58:3020–3033
67. Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26:297–302
68. Sorensen TA (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. Biol Skar 5:1–34
69. Eelbode T et al (2020) Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index. Trans Med Imaging 39:3679–3690
70. Cocianu CL, Uscatu CR (2022) Multi-scale memetic image registration. Electronics 11:278
71. Klein S, Staring M, Murphy K, Viergever MA, Pluim JP (2009) Elastix: a toolbox for intensity-based medical image registration. Trans Med Imaging 29:196–205
72. Muthukumaran D, Sivakumar M (2017) Medical image registration: a matlab based approach 2:29–34
73. Avants BB, Tustison N, Song G et al (2009) Advanced normalization tools (ANTS). Insight J 2:1–35
74. Johnson HJ, Christensen GE (2002) Consistent landmark and intensity-based image registration. Trans Med Imaging 21:450–461
75. Allasia G, Cavoretto R, De Rossi A (2012) A class of spline functions for landmark-based image registration. Math Methods Appl Sci 35:923–934
76. Habib A, Al-Ruzouq R (2005) Semi-automatic registration of multi-source satellite imagery with varying geometric resolutions. Photogramm Eng Remote Sens 71:325–332
77. Pistarelli MD, Sappa AD, Toledo R (2013) Multispectral stereo image correspondence, 217–224 (Springer)
78. Aguilera C, Barrera F, Sappa AD, Toledo R (2012) A novel SIFT-like-based approach for FIR-VS images registration. Proc Quantitative InfraRed Thermography
79. Zeng Q et al (2020) Real-time adaptive visible and infrared image registration based on morphological gradient and C_SIFT. Journal of Real-Time Image Processing 17:1103–1115
80. Zhang X et al (2021) Multimodal remote sensing image registration methods and advancements: A survey. Remote Sens 13:5128
81. Lowe DG (1999) Object recognition from local scale-invariant features, Vol. 2, 1150–1157 (IEEE)
82. Li J, Hu Q, Ai M (2019) Rift: Multi-modal image matching based on radiation-variation insensitive feature transform. Trans Image Process 29:3296–3310
83. Sedaghat A, Mokhtarzade M, Ebadi H (2011) Uniform robust scale-invariant feature matching for optical remote sensing images. Transactions on Geoscience and Remote Sensing 49:4516–4527
84. Morris NJ, Avidan S, Matusik W, Pfister H (2007) Statistics of infrared images 1–7 (IEEE)
85. Mouats T, Aouf N, Sappa AD, Aguilera C, Toledo R (2014) Multispectral stereo odometry. Trans Intell Transp Syst 16:1210–1224
86. Aguilera CA, Sappa AD, Toledo R (2015) LGHD: A feature descriptor for matching across non-linear intensity variations, 178–181 (IEEE)

87. Radhika V, Kartikeyan B, Krishna BG, Chowdhury S, Srivastava PK (2007) Robust stereo image matching for spaceborne imagery. Transactions on Geoscience and Remote Sensing 45:2993–3000
88. Wan T et al (2019) RGB-D point cloud registration via infrared and color camera. Multimedia Tools and Applications 78:33223–33246
89. Zhao D, Yang Y, Ji Z, Hu X (2014) Rapid multimodality registration based on mm-surf. Neurocomputing 131:87–97
90. Jhan J-P, Rau J-Y (2021) A generalized tool for accurate and efficient image registration of uav multi-lens multispectral cameras by n-surf matching. Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14:6353–6362
91. Zheng X, Li Z-L, Nerry F, Zhang X (2019) A new thermal infrared channel configuration for accurate land surface temperature retrieval from satellite data. Remote Sens Environ 231:111216
92. Ren H, Ye X, Liu R, Dong J, Qin Q (2017) Improving land surface temperature and emissivity retrieval from the chinese gaofen-5 satellite using a hybrid algorithm. Trans Geosci Remote Sens 56:1080–1090
93. Quan J, Zhan W, Chen Y, Wang M, Wang J (2016) Time series decomposition of remotely sensed land surface temperature and investigation of trends and seasonal variations in surface urban heat islands. Journal of Geophysical Research: Atmospheres 121:2638–2657
94. Abbasi N et al (2021) Estimating actual evapotranspiration over croplands using vegetation index methods and dynamic harvested area. Remote Sens 13:5167
95. Chen J et al (2022) A tir-visible automatic registration and geometric correction method for SDGSAT-1 thermal infrared image based on modified RIFT. Remote Sens 14:1393
96. Vijay ST, Pournami P (2018) Feature based image registration using heuristic nearest neighbour search 1–3 (IEEE)
97. Yuan Y et al (2020) Automated accurate registration method between uav image and google satellite map. Multimedia Tools Appl 79:16573–16591
98. Song X, Zheng J, Zhong F, Qin X (2018) Modeling deviations of rgb-d cameras for accurate depth map and color image registration. Multimedia Tools Appl 77:14951–14977
99. Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM 24:381–395
100. Cheng L et al (2012) Remote sensing image matching by integrating affine invariant feature extraction and RANSAC. Comput Electr Eng 38:1023–1032
101. Yang K et al (2017) Remote sensing image registration using multiple image features. Remote Sensing 9:581
102. Krishnan PT, Balasubramanian P, Jeyakumar V, Mahadevan S, Noel Joseph Raj A (2022) Intensity matching through saliency maps for thermal and visible image registration for face detection applications. The Visual Computer 1–14
103. Chen S-J, Shen H-L, Li C, Xin JH (2017) Normalized total gradient: A new measure for multispectral image registration. Trans Image Process 27:1297–1310
104. Hu H et al (2020) An artificial bee algorithm with a leading group and its application into image registration. Multimedia Tools Appl 79:14643–14669
105. Landsat 8 and srtm dataset. https://earthexplorer.usgs.gov/
106. The tufts face database. http://tdface.ece.tufts.edu/
107. Panetta K et al (2018) A comprehensive database for benchmarking imaging systems. Transactions on Pattern Analysis and Machine Intelligence 42:509–520
108. Jhan J-P, Rau J-Y, Huang C-Y (2016) Band-to-band registration and ortho-rectification of multi-lens/multispectral imagery: A case study of minimca-12 acquired by a fixed-wing uas. J Photogramm Remote Sens 114:66–77
109. Oxford dataset. http://www.robots.ox.ac.uk/vgg/research/affine/
110. Yasuma F, Mitsunaga T, Iso D, Nayar S (2008) Generalized Assorted Pixel Camera: Post-Capture Control of Resolution. Tech. rep, Dynamic Range and Spectrum
111. Mikolajczyk's dataset. http://www.robots.ox.ac.uk/vgg/research/affine
112. Group FA. Flir thermal dataset for algorithm training. https://www.flir.com/oem/adas/adas-dataset-form/
113. Xu H, Ma J, Yuan J, Le Z, Liu W (2022) RFNet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion 19679–19688
114. Debaque B et al (2022) Thermal and visible image registration using deep homography, 1–8 (IEEE)
115. Tang T, Chen T, Zhu B, Ye Y (2022) MU-NET: A multiscale unsupervised network for remote sensing image registration, Vol. 43, 537–544 (International Society for Photogrammetry and Remote Sensing)
116. Jiang X, Ma J, Xiao G, Shao Z, Guo X (2021) A review of multimodal image matching: Methods and applications. Information Fusion 73:22–71

117. Quan D et al (2022) Self-distillation feature learning network for optical and SAR image registration. Transactions on Geoscience and Remote Sensing 60
118. Parbs TJ, Koch P, Mertins A (2022) Convolutive attention for image registration 1348–1352
119. Quan D et al (2018) Deep generative matching network for optical and sar image registration 6215–6218
120. Vaswani A et al (2017) Guyon I et al (eds) Attention is all you need. (eds Guyon, I. et al.) Advances in Neural Information Processing Systems, Vol. 30 (Curran Associates, Inc.)
121. Goodfellow I et al (2014) Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger K (eds) Generative adversarial nets. (eds Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. & Weinberger, K.) Advances in Neural Information Processing Systems, Vol. 27 (Curran Associates, Inc.)
122. Li R, Zhou M, Zhang D, Yan Y, Huo Q (2023) A survey of multi-source image fusion. Multimedia Tools and Applications
123. Xu H, Ma J, Jiang J, Guo X, Ling H (2020) U2fusion: A unified unsupervised image fusion network. Transactions on Pattern Analysis and Machine Intelligence 8828:1
124. Zhang S, Zhao W, Hao X, Yang Y, Guan C (2020) A context-aware locality measure for inlier pool enrichment in stepwise image registration. Transactions on Image Processing 29:4281–4295
125. Toet A (2014) TNO image fusion dataset. https://www.figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029
126. Xu H, Ma J, Le Z, Jiang J, Guo X (2020) FusionDN: A Unified Densely Connected Network for Image Fusion. Conf on Artificial Intelligence 34:12484–12491
127. Brown M, Süsstrunk S (2011) Multi-spectral SIFT for scene category recognition, 177–184 (IEEE)
128. Wang G et al (2018) RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach 359–369 (Springer)
129. Tu Z et al (2019) RGB-T image saliency detection via collaborative graph learning. Transactions on Multimedia 22:160–173
130. Tu Z et al (2022) RGBT salient object detection: A large-scale dataset and benchmark. Transactions on Multimedia
131. Ellmauthaler A, Pagliari CL, da Silva EA, Gois JN, Neves SR (2019) A visible-light and infrared video database for performance evaluation of video/image fusion methods. Multidim Syst Signal Process 30:119–143